Paper 256-30

Decision Tree Validation: A Comprehensive Approach

Sylvain Tremblay, SAS Institute (Canada) Inc., Montreal, QC, Canada

ABSTRACT

Ever since the availability of data mining tools, decision trees have been popular because they are simple to use and easy to interpret. Producing decision trees is straightforward, but evaluating them can be a challenge. In predictive modeling, overfitting is always a danger. A good decision tree must generalize the trends in the data, and this is why the assessment phase of modeling is crucial.

SAS® Enterprise Miner™ is the SAS data mining solution. The Results window for the SAS Enterprise Miner Decision Tree node lets you visualize the tree and examine diagnostic plots and statistics. However, these components can only be viewed separately, making it harder to study the tree at varying levels of details. To extend the capabilities of the Tree node, SAS released the Tree Results Viewer. The Tree Results Viewer is a stand-alone Microsoft Windows application that significantly enhances SAS Enterprise Miner output. It can also be invoked from SAS Enterprise Miner 3.0 or later.

This paper is written for intermediate level SAS users who work with SAS Enterprise Miner 4 (SAS version 8) and does not discuss the output from SAS Enterprise Miner 5 (SAS version 9). This paper shows you how to: leverage the Tree Results Viewer in the assessment phase of decision tree modeling with SAS Enterprise Miner; interactively link the resulting tables and graphs of the tree in order to identify overfitting; evaluate the relative importance of variables in order to select the best tree.

INTRODUCTION

In a predictive modeling project, the goal is to produce a model that will generalize the patterns in the existing data and that will perform well on new data. In SAS Enterprise Miner, when the Tree node is executed, a series of trees is produced, based on previously selected Tree node parameters. By default, SAS Enterprise Miner selects the tree that maximizes the assessment measure on the Validation data set. After this occurs, the modeler has to verify that the performance of the selected tree is acceptable, and then evaluate the performance of smaller trees. For example, a smaller tree that is less complex might be more appropriate. Selecting model complexity involves a trade-off between bias and variance. An insufficiently complex model might not be flexible enough and could lead to underfitting. By contrast, an overly complex model might be too flexible and could lead to overfitting. Depending on the business context, the modeler might also have to take interpretability into account. The best tree would then be the one that has just enough flexibility (to ensure the best generalization) and, depending on the situation, the one that is interpretable by business users.

To support the quest for finding the best tree, SAS Enterprise Miner provides a number of diagnostic tools in the Results window of the Tree node. To extend the capabilities of these diagnostic tools, SAS has released the Tree Results Viewer. This viewer significantly enhances the graphical capabilities of the Enterprise Miner Tree node, providing improved presentation quality, customizable output, and enhanced printing capabilities. The goal of this paper is to show you how to use features of the Tree Results Viewer add-on to SAS Enterprise Miner when you are validating your trees.

ACCESSING, INSTALLING, AND RUNNING THE VIEWER

The Tree Results Viewer is available from the SAS Customer Support Center Web site. The viewer runs under Windows without using the SAS System and without requiring that SAS be installed. If SAS Enterprise Miner 4.0 or 4.1 is installed, then the Tree Results Viewer can be invoked from the Decision Tree node.

To access and install the viewer, follow these steps:

1. Navigate to the SAS Customer Support Center Web site at http://support.sas.com/.
2. From the left side, select **Software Downloads**.
3. Under Product & Solution Updates, select SAS System Software Products, and then select Enterprise Miner.
4. Under **For Version 8 & 6 of the SAS System**, select **Enterprise Miner Tree Results Viewer Software** http://www.sas.com/apps/demosdownloads/emtreeviewer_PROD_412_sysdep.jsp?packageID=000195.
5. Download and install the Viewer.

After installing the Tree Results Viewer, you must add the following line of code to your AUTOEXEC.SAS file (or to the start-up code of your SAS Enterprise Miner project) in order to invoke the viewer from the Tree node:

```
   %let emv4tree=1;
```

To run the Tree Results Viewer, start SAS Enterprise Miner and run the Decision Tree node. Next, right-click the Tree node and select **New view**.

GETTING STARTED

Every predictive modeling project is different. The business problem to solve, the sampling technique to be used (if any) in the construction of the model set, and the prevalence of the event in the population are a few examples of the differences that you might encounter. Whatever the situation is, you should consider using best practices before you begin to create trees and evaluate them with the Tree Results Viewer. For example, using the appropriate data partitioning and tree parameter can help to ensure that a correct assessment will result.

DATA PARTITION

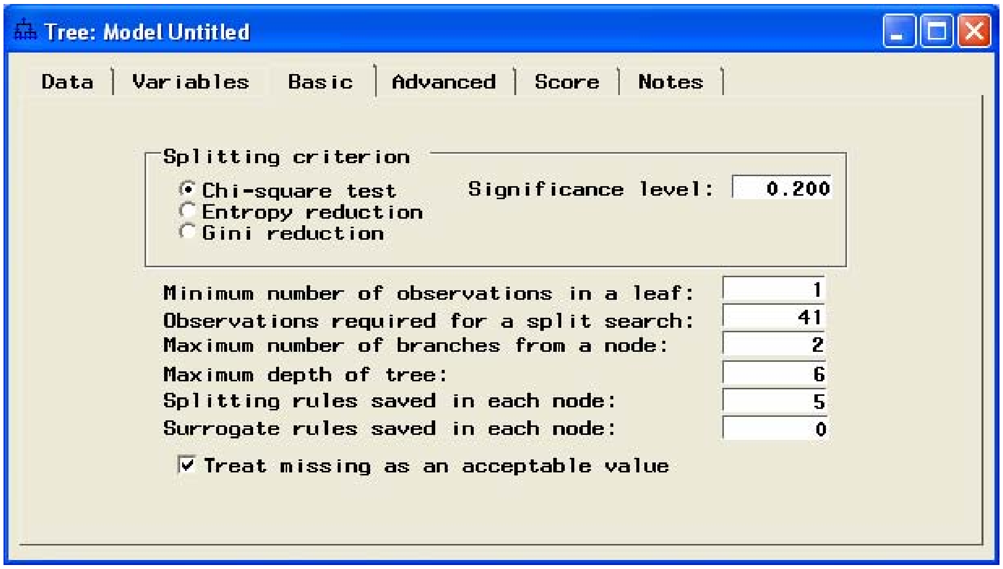
When you split your model data set into a training data set and a validation data set, be sure to create a validation data set that is large enough. A data set that is too small might lead you to erroneous conclusions when you evaluate the reliability of the tree. If you don't have enough data in your validation data set, then it will be impossible to verify a split in the training data.

TREE PARAMETERS

There are a number of tree parameters that should be set to specific values to support appropriate assessment efforts. Each is described in this section.

In the Tree node, by default the value for the **Minimum number of observations in a leaf** parameter on the Basic tab is set to 1. (See Figure 1.) This value is much too low and should always be changed because the smaller this value is, the more likely it is that the tree will overfit the training data set. If the value is too large, it is likely that the tree will underfit the training data set and miss relevant patterns in the data. The appropriate value to use depends on the context (that is, the size of the training data set); however, as a rule-of-thumb, Berry and Linoff[1] recommend setting this value to between 0.25 and 1 percent of the model set.

The **Observations required for a split search** parameter controls the depth of the tree. The value of this parameter should be set to no less than twice the value of the **Minimum number of observations in a leaf** parameter. You should also raise the value of the **Maximum depth of tree** parameter from 6 to 9 to allow a more complex tree to be grown.



Splitting criterion

● Chi-square test   Significance level: 0.200

○ Entropy reduction

○ Gini reduction

Minimum number of observations in a leaf: 1

Observations required for a split search: 41

Maximum number of branches from a node: 2

Maximum depth of tree: 6

Splitting rules saved in each node: 5

Surrogate rules saved in each node: 0

☑ Treat missing as an acceptable value

Figure 1. The Default Basic Tree Parameters in SAS Enterprise Miner

A PREDICTIVE MODELING EXAMPLE

The following example assumes that the Tree Results Viewer has been installed, and that it is available for use.

A financial institution is trying to predict which customers are most likely to default on a loan. The modeling data set is the SAS data set DMAHMEQ, which is found in the sample data sets that are installed with SAS Enterprise Miner. The location is `SASROOT\dmine\sample`. Table 1 lists the variables that are found in the DMAHMEQ data set. No sampling was used, and the file was split into a training data set and a validation data set by using a 70-30 percentage. After the data was partitioned, the Tree node was executed.

The following sections of this paper use the Tree Results Viewer to show the steps that you must take in order to select the best tree. The event that you are trying to predict (BAD=1) has an incidence of 20% in the model data set, and represents customers who have defaulted on a loan.

Variable	Role	Measurement	Type	Format	Variable Label
BAD	target	binary	num	BEST12.	Default or seriously delinquent
LOAN	input	interval	num	BEST12.	Amount of current loan request
MORTDUE	input	interval	num	BEST12.	Amount due on existing mortgage
VALUE	input	interval	num	BEST12.	Value of current property
REASON	input	binary	char	$7.00	Home improvement or debt consolidation
JOB	input	nominal	char	$6.00	Prof/exec sales mngr office self other
YOJ	input	interval	num	BEST12.	Years on current job
DEROG	input	interval	num	BEST12.	Number of major derogatory reports
DELINQ	input	interval	num	BEST12.	Number of delinquent trade lines
CLAGE	input	interval	num	BEST12.	Age of oldest trade line in months
NINQ	input	interval	num	BEST12.	Number of recent credit inquiries
CLNO	input	interval	num	BEST12.	Number of trade (credit) lines
DEBTINC	input	interval	num	BEST12.	Debt to income ratio

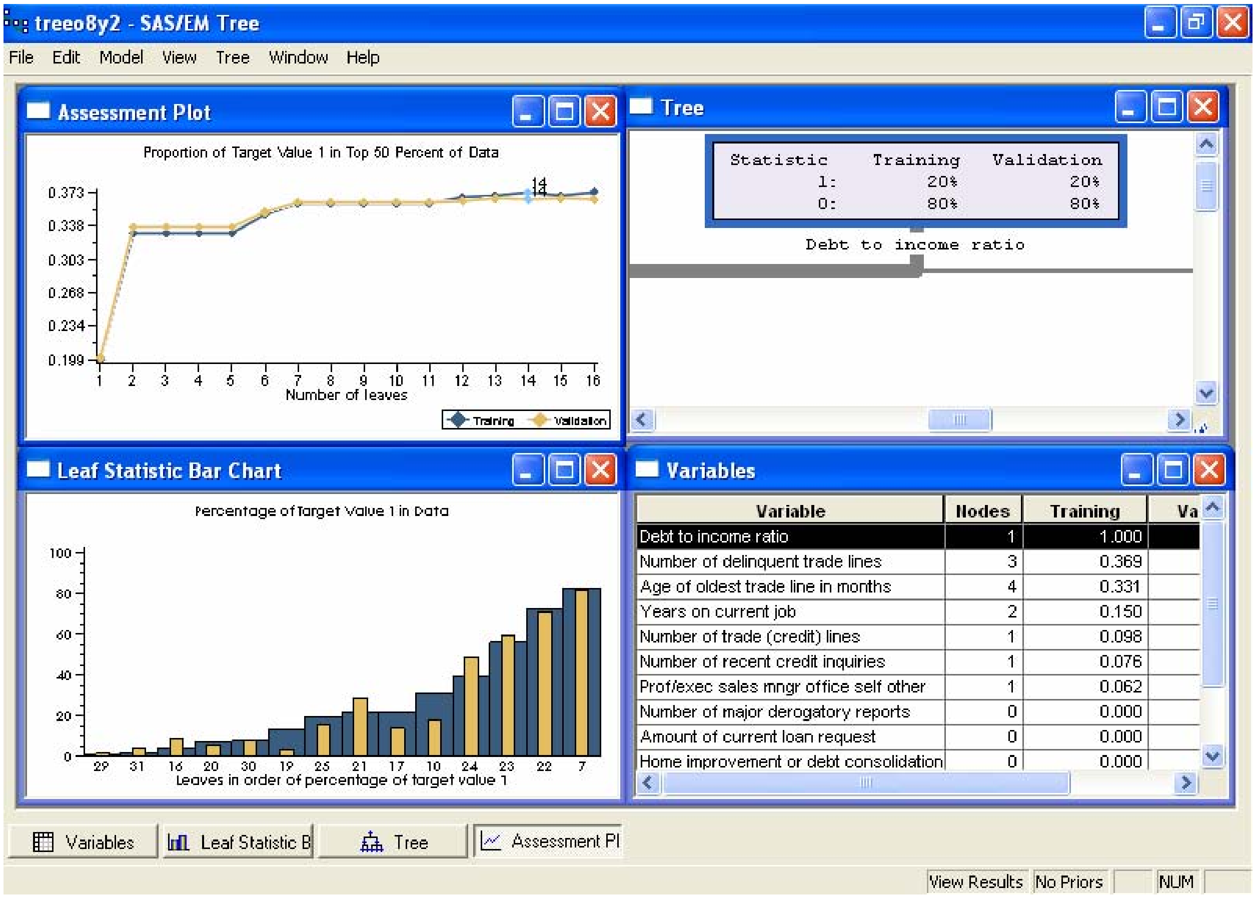
Table 1. The Variables in the DMAHMEQ Data Set

THE TREE RESULTS VIEWER INTERFACE

After running the tree, right-click the Tree node and select **New View** in order to invoke the Tree Results Viewer. Figure 2 shows the viewer’s interface. The viewer displays a maximum of 14 tables and graphs in separate windows. Each window can be independently arranged, and related windows are dynamically linked together.

For example, clicking a variable, node, or subtree in one window automatically updates and selects the corresponding items in the other window. All views can be copied or printed. These features enable you to easily capture components of your tree and, if necessary, place them in a Microsoft Word document. As you will see in the following sections, these features make your job as modeler much easier.

Also, if you set prior probabilities in your SAS Enterprise Miner flow, you can apply these priors through the Edit menu of the Tree Results Viewer, and the windows of the viewer will be updated accordingly.



```
Statistic     Training   Validation
       1:          20%          20%
       0:          80%          80%
```

Debt to income ratio

Variable	Nodes	Training	Va
Debt to income ratio	1	1.000	
Number of delinquent trade lines	3	0.369	
Age of oldest trade line in months	4	0.331	
Years on current job	2	0.150	
Number of trade (credit) lines	1	0.098	
Number of recent credit inquiries	1	0.076	
Prof/exec sales mngr office self other	1	0.062	
Number of major derogatory reports	0	0.000	
Amount of current loan request	0	0.000	
Home improvement or debt consolidation	0	0.000	

Figure 2. The Results Tree Viewer Interface

THE ASSESSMENT PLOT

The first thing to review in the Tree Results Viewer is the Assessment Plot, which shows tree evaluation information. Trees are evaluated using the proportion of cases that are correctly classified in the top 50% of the training data (as set in the SAS Enterprise Miner Tree node). For each tree size, a tree that correctly classifies the most training cases is selected to represent that size. The selected tree is evaluated again with validation cases. In this plot, it's best to identify the following conditions:

- whether the line for the training data is progressing
- the extent to which the validation data confirms the progress
- how much progress seems achievable while using the smallest tree possible

In Figure 3, the line for the training data is progressing as the number of leaves increases. The validation data confirms this progress, but you can see that, starting from the tree that has 12 leaves, there is a slight difference (although not enough to make the tree unreliable). The tree that was chosen by SAS Enterprise Miner as the best tree to use was the one with 14 leaves because that tree maximizes the assessment value on the training data set. In terms of assessment, the chosen tree has a lift of 4 for the first decile and a lift of 1.8 (0.36/0.20) for the top 5 deciles, so it does a decent job. However, could a less complex tree perform as well in terms of assessment while also being more reliable? Let's investigate this further, but before doing that, examine the tree that has 14 leaves in terms of reliability. You might discover why the progression differs between the training and the validation data sets (in terms of performance).

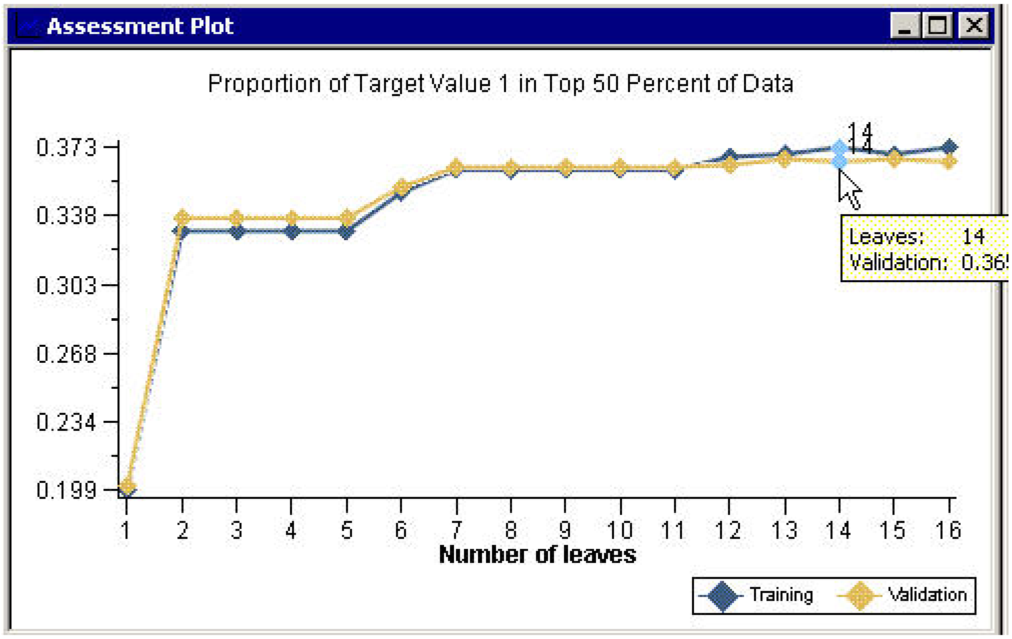


Figure 3. The Assessment Plot in the Tree Viewer

THE LEAF STATISTIC BAR CHART

Let's determine if the tree that has 14 leaves is reliable. To answer that question by using SAS Enterprise Miner, you would have to look throughout the whole tree for unreliable leaves. For large trees, this takes time because it is impossible to see the tree globally and find faulty leaves rapidly. However, using the Tree Results Viewer, this task becomes very easy. Select the tree that has 14 leaves and keep it selected in the Assessment Plot (that is, don't click a smaller tree). Then select the Leaf Statistic Bar Chart (see Figure 4).

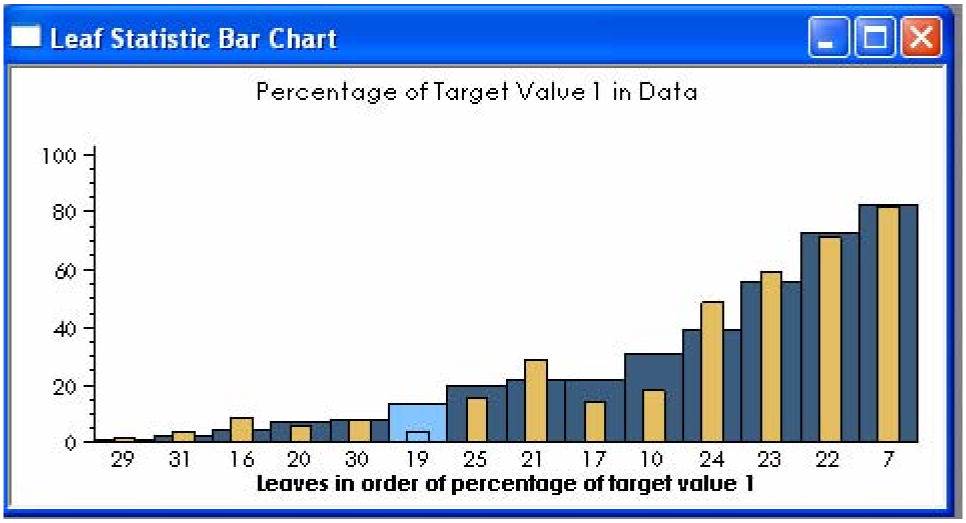
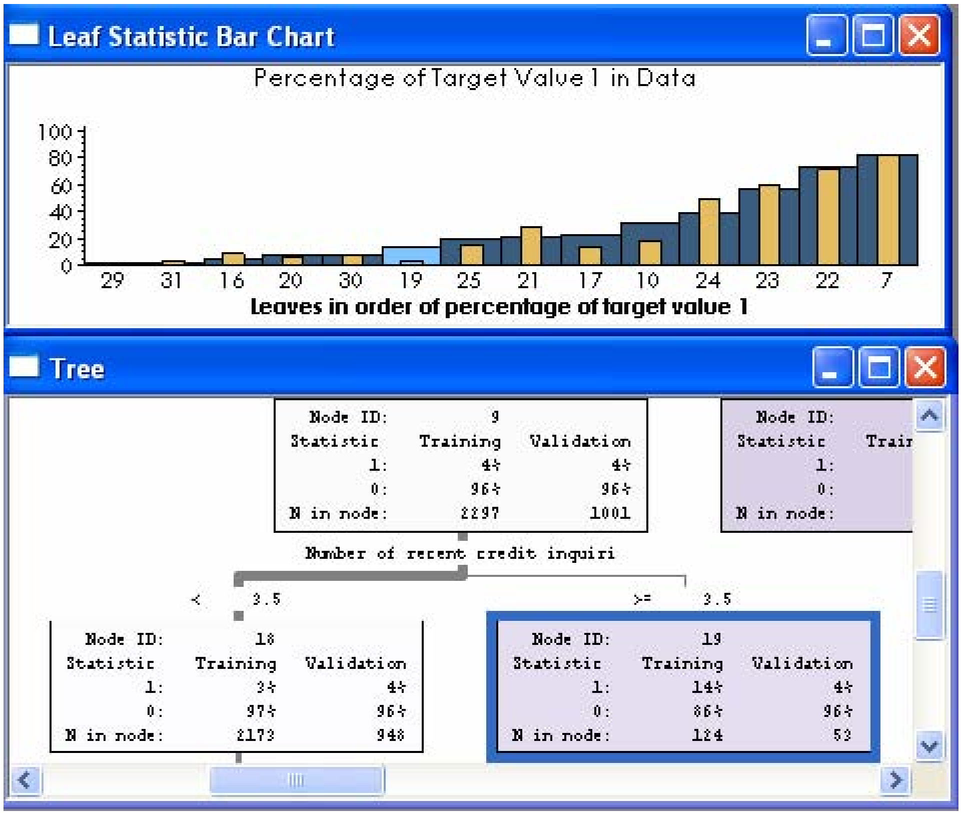

Leaf Statistic Bar Chart
Percentage of Target Value 1 in Data
100
80
60
40
20
0
29 31 16 20 30 19 25 21 17 10 24 23 22 7
Leaves in order of percentage of target value 1

Figure 4. The Leaf Statistic Bar Chart Window

The bars represent the proportion of events (BAD=1) for each leaf for the training data set (blue bars) and the validation data set (yellow bars). The bars are sorted in order of increasing values of the proportion of events in the training data. If the tree were reliable, a similar increase in bar heights would be seen in the validation data, except in some of the small leaves where there is not enough data to make reliable estimates. Overall, you do see this increase, but some of the leaves (19, 17, 10, and 24) look unreliable. As an example, the validation data for Leaf 19 does not confirm what you see in the training data. To investigate this leaf, click the bar for Leaf 19 in order to highlight it and its associated leaf in the Tree window. By viewing these two windows at the same time, as shown in Figure 5, you can see which leaf to investigate.


Leaf Statistic Bar Chart
Percentage of Target Value 1 in Data
100
80
60
40
20
0
29
31
16
20
30
19
25
21
17
10
24
23
22
7
Leaves in order of percentage of target value 1
Tree

```
Node ID:          9
Statistic    Training   Validation
       1:          4%           4%
       0:         96%          96%
N in node:      2297         1001

Number of recent credit inquiri

<    3.5                                   >=    3.5

Node ID:         18               Node ID:         19
Statistic   Training  Validation  Statistic   Training  Validation
       1:         3%          4%         1:        14%          4%
       0:        97%         96%         0:        86%         96%
N in node:      2173         948  N in node:       124          53

Node ID:
Statistic    Trair
       1:
       0:
N in node:
```

Figure 5. The Link between the Leaf Statistic Bar Chart Window and the Tree Window

In Node 9 (the parent of Leaf 19), the proportion of defaulters is 4% in both the training and the validation data sets. This node is then split into Node 18 and Leaf 19. In Leaf 19, the proportion of event is higher (14%) for the training data set but stays the same (4%) for the validation data set. Additionally, there are only 53 observations in the Validation data set. This makes the leaf unreliable.

THE TREE WINDOW

The Tree window in the Tree Results Viewer is very flexible. As is the case with SAS Enterprise Miner, you can decide which statistics to display, which color scheme to use, and so on. However, unlike SAS Enterprise Miner, you can zoom in and zoom out in order to view the tree at the most appropriate degree of detail. Therefore, printing the tree is simple. You can also copy the tree (zoomed view or regular view) and paste it into a Microsoft Word document. Figure 6 illustrates this flexibility.

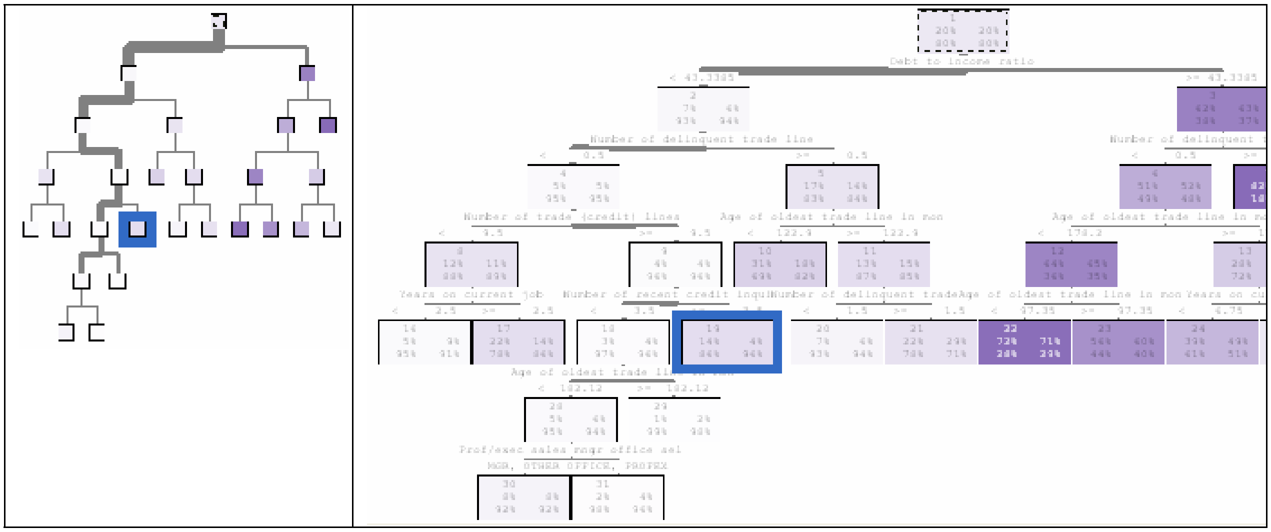
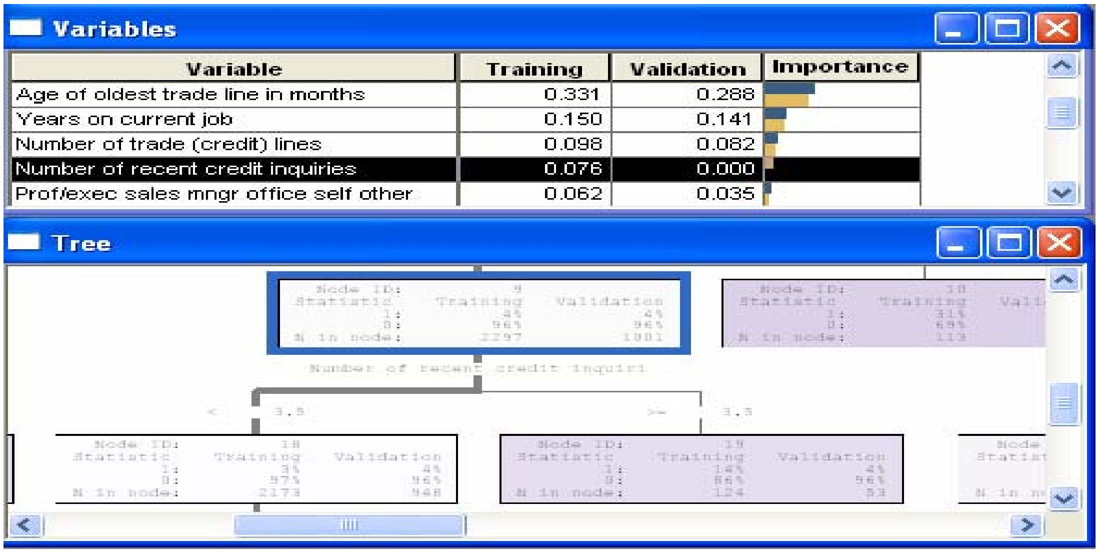

Debt to income ratio
< 43.3385
>= 43.3385
Number of delinquent trade line
< 0.5
>= 0.5
Number of trade (credit) lines
Age of oldest trade line in mon
Years on current job
Number of recent credit inqu
Number of delinquent trade
Age of oldest trade line in mon
Prof/exec sales mngr office sel
MGR, OTHER OFFICE, PROFEX

Figure 6. A View of the Tree at Various Degrees of Detail

THE VARIABLES IMPORTANCE TABLE

In addition to the Leaf Statistic Bar Chart window, there is another window that can help you find unreliable leaves, that is, the Variables window. The Variables window displays a table that lists the variables in the order of their importance in the tree. The last column in the table uses horizontal bars to compare the training and validation estimates of the importance of the variables. The top bars represent the training data set, and the bottom bars represent the validation data set. Where the validation bar is noticeably shorter than the training bar, a deceptive split might be the cause. As shown in Figure 7, the variable **Number of recent credit inquiries** has some importance on the training side but no importance on the validation side. If you select that variable in the Variables window, it will highlight the node that is responsible for the deceptive split within the Tree window. Node 9 is the culprit and is also the same node that was previously identified through the Leaf Statistic Bar Chart.



Variable	Training	Validation	Importance
Age of oldest trade line in months	0.331	0.288	
Years on current job	0.150	0.141	
Number of trade (credit) lines	0.098	0.082	
Number of recent credit inquiries	0.076	0.000	
Prof/exec sales mngr office self other	0.062	0.035	

```
Node ID:            9
Statistic    Training   Validation
       1:          4%           4%
       0:         96%          96%
N in node:       2297         1001

Node ID:           10
Statistic    Training   Vali
       1:         31%
       0:         69%
N in node:        113

Number of recent credit inquiri
<   3.5                    >=   3.5

Node ID:           18
Statistic    Training   Validation
       1:          3%           4%
       0:         97%          96%
N in node:       2173          948

Node ID:           19
Statistic    Training   Validation
       1:         14%           4%
       0:         86%          96%
N in node:        124           53
```

Figure 7. Viewing Variables Window and Tree Window Concurrently

RESULTS OF THE ANALYSIS

After finishing the analysis, you would most likely conclude that the tree that has 14 leaves performs well in terms of the assessment measure, but that the tree has too many unreliable leaves. A less complex, but more reliable, tree might be more appropriate. Moving back to the Assessment Plot, you can see that the tree with 7 leaves is much simpler and still performs well. To verify that this tree is reliable, you must perform another analysis using the techniques that were described previously.

CONCLUSION

Decision trees are easy to create, but evaluating them requires both the ability to visualize the trees at different levels of detail and the flexibility to link the trees to various diagnostic charts and plots. The Tree Results Viewer complements SAS Enterprise Miner 4 (under SAS version 8) by giving the user this needed ability and flexibility. In particular, the Tree Results Viewer gives you more control over the output of the tree. When you use this viewer, the dynamic links between the diagnostic charts and the tree make it easier for you to find the best tree in a vast forest of candidates. Additionally, the viewer makes it easy to print the tree.

Note: The Tree Results Viewer has been improved even further and is now fully integrated into SAS Enterprise Miner 5, which is compatible with SAS®9. Under this latest release of SAS Enterprise Miner, you can use this viewer to interactively train a decision tree. The latest version of the Tree Results Viewer also provides several advanced visualizations.

REFERENCES

1. Berry, M.J.A., and G. Linoff, G. 1999. Mastering Data Mining: The Art and Science of Customer Relationship Management. New York: John Wiley & Sons, Inc.

2. Barlow, T., and P. Neville. 2001. “Case Study: Visualization for Decision Tree Analysis in Data Mining.” *Proceedings of the IEEE Symposium on Information Visualization 2001*, San Diego, CA: 2001, 149-152.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Sylvain Tremblay
SAS Institute (Canada) Inc.
1000 Sherbrooke Street West, Suite 2100
Montreal, Quebec, Canada, H3A 3G4
Tel (514) 395-4092
Fax (514) 395-8962
Sylvain.Tremblay@sas.com
http://support.sas.com/training/canada/